Just how difficult can it be counting up R&D funding for emerging technologies (and is tech mining with proxy measures going to be any better?)

Article (Published Version)

Technology Analysis & Strategic Management

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/ctas20

Just how difficult can it be counting up R&D funding for emerging technologies (and is tech mining with proxy measures going to be any better)?

Michael M. Hopkins [a] & Josh Siepel [a]

[a] SPRU, Freeman Centre, University of Sussex , UK

Published online: 18 Jun 2013.

To cite this article: Michael M. Hopkins & Josh Siepel (2013) Just how difficult can it be counting up R&D funding for emerging technologies (and is tech mining with proxy measures going to be any better)?, Technology Analysis & Strategic Management, 25:6, 655-685, DOI: 10.1080/09537325.2013.801950

To link to this article: http://dx.doi.org/10.1080/09537325.2013.801950

Technology Analysis & Strategic Management, 2013
Vol. 25, No. 6, 655–685, http://dx.doi.org/10.1080/09537325.2013.801950

Just how difficult can it be counting up R&D funding for emerging technologies (and is tech mining with proxy measures going to be any better)?

Michael M. Hopkins* and Josh Siepel

SPRU, Freeman Centre, University of Sussex, UK

Decision makers considering policy or strategy related to the development of emerging technologies expect high quality data on the support for different technological options in order to track trends and allocate resources. A natural starting point would be R&D funding statistics. This paper explores the limitations of such aggregated data in relation to the substance and quantification of funding for emerging technologies. Using biotechnology as an illustrative case, we test the utility of a novel taxonomy to demonstrate the endemic weaknesses in the availability and quality of data from public and private sources. Using the same taxonomy, we consider the extent to which tech-mining presents an alternative, or potentially complementary, way to determine support for emerging technologies using proxy measures such as patents and scientific publications. We find that using proxy measures provides additional visibility of technological emergence and suggest these are a useful complement to financial data.

Keywords: science and technology and innovation policy studies; technology and innovation studies; science and technology indicators; emerging technologies; data mining; social shaping of technology

1. Introduction

Emerging technologies, when combined with innovative entrepreneurs and complementary institutional changes, have the potential to generate super-profits and economic growth (Freeman and Louçã 2001). Harnessing such technologies is a key objective for government policy makers and industrialists alike, with the means to evaluate related activity and progress increasingly valued (Martin 1996; Patel 2012). In the present economically challenging times, as funds invested in R&D are restricted, the ability to make important decisions that will affect the viability of particular technological options becomes even more important.

Yet decisions about the funding of technological options are not made in a vacuum, as emerging technologies are socially shaped and are selected over other potential options (Bijker 1995;

*Corresponding author. Email: m.m.hopkins@sussex.ac.uk

Smith, Voß, and Grin 2010). The outcomes of this struggle often reflect the agendas of powerful actor groups with vested interests (Jacobsson and Johnson 2000; Verbong and Geels 2007). The thinking of these groups, their norms and routines thus constrain scientific and technological paradigms (Dosi 1982). For example, the emerging literature on 'pharmaceuticalisation' examines how heavy private investment in pharmaceutical R&D and marketing has led to an assumed societal focus on pharmaceutical approaches to mitigating disease, at the expense of non-pharmaceutical approaches that are either underutilised or ignored (Abraham 2010; Williams, Martin, and Gabe 2011).

It follows that when policymakers, firms and academics review the funding of a range of distinct technological options they will often be dependent upon information about these options derived in a highly politicised context, with R&D funding data not being exempt from influence even if this manifests as a lack of will to generate accurate data. The continued absence of suitably complete data (i.e. spanning public and private sectors, national boundaries, different technological fields) in economically important areas such as energy and bioscience has recently been noted in unrelated fields (Anon 2010; House of Lords 2010). This is despite a notable emphasis in some countries such as the USA and UK on developing systems to record investments and related impacts by public sector and charitable research funders.[1] In the absence of robust, complete and objective data on funding, not only is there a lack of accountability but also the potential for loss of national capacity and opportunities if clear figures for R&D investment are not available to policymakers. For example, in 2010 a UK parliamentary Committee considering the prioritisation of national public R&D spending found that 'aggregated information on how much is spent on particular aspects of research and development within key sectors was not readily available; as a result, it was difficult to identify the reasons why resources were distributed as they were' (para. 28).[2] The problem is not limited to the UK. When Texas state Senator Steve Ogden sought to ban stem cell research in the state in 2009, a lack of data on funding stymied his attempts, as it was unclear how extensive the impact of legislative change would be on Texan institutions (Matthews and Rowland 2011).

This paper addresses the following questions: (i) Why is relevant R&D funding data often in fact misleading or unavailable? (ii) How best can data be obtained to inform decision making to support the development of emerging technological options in science-intensive areas of technological change?

In addressing these questions we note how this focus poses distinct methodological challenges to those described in prior work on measuring innovative activity or research quality (previously reviewed by Patel (2012) and Martin (1996)).

Regarding the first question, we explore the biases and weaknesses in a wide range of data used to analyse the R&D landscape, using emerging biotechnologies as exemplars. We argue that the types of funding data available are not as representative or comparable as is necessary for the purposes of tracking technologies. This is a distinct challenge from other perhaps more easily bounded subject matter such as reporting on academic disciplines, or national activity). We caution against the assumption that data on R&D funding conforms to the same apparently more objective characteristics of data studied by scientists within their research fields, and discuss here why attempts to gather more complete funding data will likely not resolve the problem. In this way the paper builds on work by NESTA (2006) that discussed the flaws and weaknesses of singular national statistics, and extends a critique made by one of this paper's authors in a recent discussion paper, also on the funding of emerging technologies, for the Nuffield Council on Bioethics (see Nuffield Council on Bioethics 2012).

On the second question, of alternatives to using R&D funding data, we discuss the utility of tech mining approaches using proxy measures (Cozzens et al. 2010) for tracking emerging technologies.

The paper's theoretically driven critique suggests why proxy measures of technological emergence, despite their drawbacks, will often be insightful for the purposes of tracking technological options, potentially to a greater extent than more 'direct' measures of innovation such as funding figures.

However, we illustrate proxy measures are diverse in characteristics and no individual measure will be universally applicable. They must be carefully chosen to be suitable for the fields to be studied. Strengths and weaknesses of both direct measures of R&D funding and of specific proxy measures such as patents and publications are explored using a novel taxonomic framework that builds on conceptualisation of classification issues related to substance and quantification in Science and Technology Studies (STS), which we set out in Section 2, below. Section 3 discusses the methods we use to undertake the analysis. Section 4 applies the taxonomy to the problem of tracking R&D funding in biotechnology, focusing mainly on developed countries. Having demonstrated the difficulties in collecting R&D funding data on emerging technologies, Section 5 explores some of the benefits and limitations of alternative proxy measures, which are illustrated in Section 6 through a case study of genomic technology. Section 7 discusses our findings and in Section 8 we draw our conclusions.

2. Social shaping of substance and quantification in R&D statistics

In this section we describe how an established STS literature on the socially mediated processes of classifications can be used to provide a helpful framework for understanding a novel problem, namely establishing investment in emerging technologies and differential investment in distinct technological options. The challenge here is to understand how collecting and organising information turns otherwise formless data into something useful (Headrick 2000). As an illustration of why this is worth unpacking, even an apparently simple statistical observation such as 'Germany spends more on R&D than the European average' cannot be fully understood without positioning elements of the statement in at least seven of the ten possible Aristotelian classifications – substance, quantity, where, when, action, having, relations (Aristotle 1995), making this apparent fact a highly complex ontological construct.[3] Here we focus mainly on the challenges of classifying substance and quantity, which we suggest raise more particular concerns in relation to the study of technologies, while the positioning of investments in time and space, for example being more generic issues, but not necessarily less problematic.

Substance

In referring to substance we mean the type of entity, being, or phenomenon that is to be defined and classified. This is a most fundamental starting point. Indeed Power (2005, 586) suggests 'without a system of concepts and taxonomy any practice of intervention is blind, disorganised and of questionable legitimacy'.

Substances rarely exist in isolation in their natural state and so must be delineated from their context through the process of definition. Definitions are therefore artificial ontological constructs that are valid only within certain communities. Definitions are often devised through contentious processes which may be hidden to those beyond (cognitively, spatially, or temporally) the immediate debate (Power 2004, 2005). Different communities will have different classifications that suit their individual needs as parties make formal or informal attempts to arrange their classification to reflect those issues (Bowker and Star 2000). Thus classification systems may have different

scope, and be judged to be too broad or narrow or even irrelevant, by those in other communities. As a substances are discussed across communities they may become 'boundary objects', which:

> inhabit several communities of practice and satisfy the informational requirements of each of them. Boundary objects are thus both plastic enough to adapt to local needs and constraints of the several parties employing them, yet robust enough to maintain common identity across sites. They are weakly structured in common use and become strongly structured in individual site use. (Star and Griesemer 1989)

Amid different classification systems, boundary objects provide ideas that link disparate areas of practice. Yet between different systems those broadly understood terms may be the main links, with subsidiary classifications unresolved. This process of classification underpins our broad view of the world, from medicine (Bowker and Star 2000) to economics (Hicks 2011). Once created, classifications may be used for political purposes such as resource allocation, thus unlike natural phenomena which have an existence outside of human agency – what Searle (1995) refers to as 'brute facts', classifications create 'institutional facts' (ibid.) and so classification may be performative (Power 2005): in other words, it can bring substances into being. Furthermore politically popular options will attract re-framing of activities. For example, Calvert (2006) observes how R&D projects can shift between definitions of 'basic' or 'applied' research or from one subject to another according to research funding policy.

It follows from the reasoning above that the definition of emerging technological options – the *Substance* (*S*) we are interested in here, will be problematic because of: (S.1) *novelty* – they are new, and therefore may not have been the subject of detailed ontological debate, or the results of such debate may not yet be diffused. Perhaps they may not be perceived as important enough to classify/ collect data on; (S.2) *plurality* – of communities and therefore distinct definitions (either overlapping or orthogonal) based on differing perspectives of what should be categorised and how, thus inhibiting comparison. (S.3) *boundary objects* – where the same substance is recognised by different communities in similar and distinct ways, promoting misinterpretations (S.4) *indistinctness* – technologies are still emerging, fusing or evolving and it may not be clear how they are distinct from prior substance; (S.5) *performativity* – technological options may be fed or starved by classification systems when these are coupled to resource allocation with categories created to reflect the interests of those with power.

Quantification

Quantification itself is an outcome of classifications: once things are classified they can then be counted (Power 2004). Within the sciences being able to count things is a highly desirable outcome – Porter (1992) characterises the desire for quantification as the 'accounting ideal' – that is, for the natural world to be clearly defined as a balance sheet. Yet the process of arriving at a numerical result is not trivial, but instead heavily contested – given a political tendency for 'trust in numbers' (Porter 1996) the process of making the end result becomes crucial. For instance, the process of determining what gets counted and what does not is therefore inherently political, and the determination of the 'rules' by which data is collected can have meaningful impacts on the management decisions taken with the data, as seen in a discussion of the calculation of Canadian crime statistics (Haggerty 2001).

The following summary draws exclusively on recent studies of performance measures in management (Power 2004) and financial risk in banking (Power 2005) to provide an extensive (but not

comprehensive) account of why *Quantification* (Q) of data is often problematic: (Q.1) *Method* – the ability to collect and measure data may not exist, even if it is possible to conceive a classification system (for example, how many black holes are there in the universe?); (Q.2) *Quanta* – the unit of measurement must be agreed and data collected in that mode e.g. which currency to count in, how to relate historical and current prices; (Q.3) *Norms* – where different methods of counting may be employed by different agents, for example, what thresholds should be set for minimum values worth collecting? (Q.4) *Jurisdiction* – data sources are dispersed such that they are the property of multiple agents or research participants and subject to access restrictions, or associated costs, making the output of collection uneven. (Q.5) *Incentives* – agents or research subjects may be likely to materially benefit or lose out from under or over reporting of data thus distorting search results; (Q.6) *Judgement* – ambiguity over which of two or more classifications to allocate an instantiation opens the way to subjectivity and potentially bias; (Q.7) *Confidentiality* – some data may be unavailable owing to commercial concerns or contractual obligations; (Q8) *Loss* – the subject of study may have decayed, or may no longer exist, may be hidden or lost; (Q.9) *Audit* – quantification processes may yield to replication, transparency of method, and therefore establishment of best practice, monitoring and trust, and these may all influence the effectiveness of data collection undertaken as part of an established routine.

Additional to prior findings reviewed above, we propose two additional factors. First, stemming from the observation that technologies are not immutable, we might expect (Q.10) *transformation* – of substance and/or classification. Change over time means that what begins as funding for one field may yield results that can be classified in another at a later date. Equally classification systems will transform as well thus categories will split or fuse. Finally, (Q.11) *Cost* – it must be recognised that in many cases data is not valued sufficiently for the costs of collection to be met.

The taxonomised issues above are set out in Table 1 which groups these 16 issues into five wider categories, two relating to Substance (Definition, Emergence) and three relating to quantification (Counting processes; Data gathering; Data decay). We now move to explore the extent to which direct and indirect measures of technological options suffer from these problems.

3. Method

In the following section we demonstrate how the above issues arise in attempts to track R&D funding in technological options using the case of emerging biotechnologies. The empirical work is based on review of an illustrative range of the secondary sources where extensive accounts of activity in biomedical R&D are provided focusing mainly on recent and widely available work produced the UK, EU and USA. Additional insights were gained from four interviews with authors of four reports that have attempted to collate R&D funding in the field, based on a convenience sample. This is sufficient because the primary aim of the above empirical strategy is purely to illustrate that factors S1–5 and Q1–11 have impinged on attempts to classify substance and quantity in relation to emerging technologies, and thereby to validate the utility taxonomy presented in the previous section.

In Section 6 we draw on prior published research to illustrate the strengths and weaknesses of Tech-Mining using proxy measures in relation to the same factors mentioned above. An example is provided drawing on prior research by one of the authors, together with colleagues, on the emergence and use of genomic technology in biomedicine. The methods used for these studies are referred to in the text where essential to demonstrate key points but due to space limitations here, full details may be found in the original publications only.

Table 1. Factors affecting data codification and comparison between R&D funding and patent/publication proxies

Category	Facet of codification	R&D investment figures	Patents	Publications
Substance: Definition	Plurality	Different providers of data use different definitions. R&D figures generated by diverse private organisations make it difficult to account for specific emerging technologies; public R&D figures are missing data and comparability between funders is often not possible where definitions differ	Flexible, classification system based on key words and classes (Strumsky et al. 2012)	Flexible classifications available using author, institution, keyword etc. (Rafols et al. 2012)
	Indistinctness	Difficult to distinguish the contribution of or to a new technology when outputs can be produced using older technologies	Citations allow the influence of distinct cognitive approaches to be mapped, so long as the technology is patentable and not subject to other forms of appropriation (Hall, Jaffe, and Trajtenberg 2001)	Citations allow the influence of distinct cognitive approaches to be mapped but data may be incomplete owing to problems with matching (Moed 2002)
	Boundary objects	Lack of a common definition across boundaries makes 'biotechnology' commonly understood but with different meanings	Classification systems make definition explicit but may not fully represent characteristics of technologies (Benner and Waldfogel 2008)	Classifications make definition explicit, though citation process is socially mediated (Nicholaisen 2008)

Substance: Emergence	Novelty	Research on discrete new areas is difficult to observe as it is not recognised. In some cases 'bootlegging' means work carried out may not reflect stated purposes of funding for new projects in unsanctioned areas of investigation (Augsdorfer 1996)	New technologies and patent categories emerge following regular reviews but terminology may change over time	New journals are created and categories are altered, but these may be delayed
	Peformativity	Tendency for categories to reflect areas of interest reflects back on demand for funding	Many patent citations added by examiners, introducing noise (Alcacer and Gittelman 2006)	Self-citation, enforced citation and other game-playing impact on publication behaviour (Bornmann 2011)
Quantification: Processes of counting	Quanta	Varies with definition: for instance 'biotechnology', 'firm', 'funder', etc.	Clearly defined but not representative of all innovation: some innovations not patentable, others not appropriable with patents (Hall, Jaffe, and Trajtenberg 2001)	Clearly defined but not necessarily equal owing to uneven definitions, 'salami slicing' and the Matthew effect (Bornmann 2011)
	Method	Difficult to identify the full range of funders and their influences – tendency to assume observed research funding is representative of the global level, which is misleading	Large datasets are already collected with all active organisations represented in patent database universes at least in major markets such as developed countries (e.g. patstat, USPTO)	Large datasets are already collected with many active organisations globally – although subject to language biases; logistical issues posed by download limits, citation matching, etc.

(*Continued*).

Table 1. Continued.

Category	Facet of codification	R&D investment figures	Patents	Publications
	Norms/thresholds	Different countries/organisations use different definitions of biotechnology but also count activities of a particular scale (e.g. only dedicated biotech firms or firms whose major activity is biotechnology) and so calculate their figures differently	Criteria for granting patents may exclude some innovations	Threshold for publication in different journals varies
	Judgement	Researchers must make a subjective judgement about the point at which a firm may have sufficient 'biotechnology' activities to be considered as a biotechnology firm. Where many are involved in this process, differences in judgement will make data less consistent	Researcher is reliant on indexes, and externally generated and awarded classifications; patent examiners assign many citations (Tan and Roberts 2010)	Researcher is reliant on publication indexes, as well as and externally generated classifications
	Audit	Where data is collected regularly, routines for consistent can develop. But different auditing systems (RePORT, RCUK's gateway) do not guarantee compatibility of data between systems	Internal review processes for patent authorities ensure validity of work within a jurisdiction.	Peer review system has its own biases (Hojat, Gonella, and Caelleigh 2003), as do citations and journal impact rankings (Glanzel and Moed 2002)

Quantification: Data collection	Incentives	Respondents not incentivesed to share data; international partners potentially not incentivised to collaborate	Financial incentives to patenting, where innovation is patentable and appropriable	Incentives to publish vary, with different incentives for publishing in indexed journals (Veugelers 2005); incentives may lead to over/under-stating private sector research)
	Jurisdiction	National funding agencies and other bodies may restrict access to data, making comparison difficult	Different norms for patent procedures and citations between US and EU may make direct comparison difficult	Non-English publication language biases may understate research across countries
	Confidentiality	Proprietary data and commercial confidentiality make data collection difficult	Most data public, where patents exist	Most data available by subscription, though some content remains behind paywalls
	Cost	Highly intensive of time and effort	Low cost or free; additional time required for cleaning	Low cost or free; Generally limited to appropriate subscriptions
Quantification: Data decay	Transformation	Purposes of technologies change over time, meaning that *ex ante* and *ex post* funding declarations differ	Potentially difficult to track over time using traditional bibliographic techniques	Potentially difficult to track over time using traditional bibliographic techniques
	Loss	Perceptions of investments change over time such that projects funded and observed *ex ante* are, *ex post*, not visible and funding ascribed to successful projects	Data is quantified and long-run, but may be 'lost' with changes in language use, keywords and category changes	Data is quantified and long-run but may be 'lost' with changes in language use, keywords, and category changes

Summary: R&D Investment vs proxy figures.

4. The classification of R&D funding in emerging biotechnologies

4.1. *Defining the substance for data collection*

The taxonomy described in Section 2 presents a number of key challenges to the description and quantification of the emergence of new technological options. Out of the issues identified there (summarised in Table 1), we may identify two sets of challenges of identifying the substance of new and emerging technologies: definition and emergence.

4.1.1. *Definition*

Even among existing technological categories, there are several issues that make the establishment of cleanly cut categories fundamentally difficult (OECD 2009). This is even the case for the OECD, which since Chris Freeman's initial efforts in the 1960s have published six editions of the Frascati Manual, dedicated to the accurate collection of international R&D statistics.[4] The plurality and indistinctness inherent in the categories used to classify technologies, and development of boundary objects common in discussions of innovation, complicate the establishment of reliable grounds for comparison. Inevitably in describing when these issues arise we will encounter in passing other related problems to do with quantification (such as judgement) however we will revisit these later in our account.

In the context of our illustrative analysis on emerging biotechnologies, it is important to highlight that even the basic definition of key terms such as life sciences industry or biotech firm are not necessarily agreed, or clear (BIS 2010; OECD 2009). For example the journal *Nature Biotechnology* has a long running annual global survey of the stock-market listed biotech firms but notes 'as the industry has grown and changed so has our definition of what constitutes a biotech company and our methods for gathering data' (Huggett et al. 2011, 585). In part this is because data gathering complexity has reached the scale that *Nature Biotechnology* now has to rely on the international consultancy, Ernst and Young, whose own definition is distinct.

The OECD has also tried to develop a global survey to support a detailed statistical resource focused on 'biotechnology' for which it has established an internationally used definition:

> The application of science and technology to living organisms, as well as parts, products and models thereof, to alter living or non-living materials for the production of knowledge, goods and services. (OECD 2005)

However the OECD rely on member states to undertake national surveys and these often use distinct definitions (OECD 2009). Furthermore broad definitions such as the above are open to judgement during attempts at quantification. For example, application of science and technology to living organisms, might be interpreted as developing purely mechanical devices (e.g. stents) and applying these to living organisms. The OECD therefore recommends that single definitions be accompanied by lists of technical approaches that would fall within the definition (OECD 2009, 9). Thus plurality of definitions disrupts comparability unless carefully considered in co-ordinated data collection efforts.

When the OECD sought to collect comparable statistics on biotech firm numbers in 26 countries, surveys and data gathering was undertaken by agents in different jurisdictions, the term 'biotechnology' became a boundary object as there different ways to classify the firms as Figure 1 illustrates. Also fundamental is distinguishing between biotechnology and traditional technologies used in pharmaceutical R&D, one of the largest users of biotechnology: What is the contribution of biotechnologies, for example tools like gene cloning, to development of pharmaceuticals? The

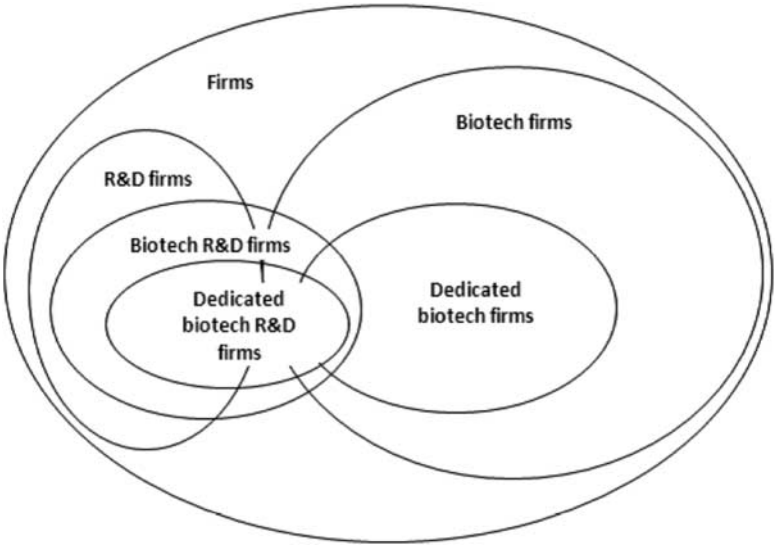

Figure 1. Definitions of the biotechnology industry.
Source: OECD (2009).

indistinctness between categories of 'biotechnology' and other health-related R&D makes this more difficult to measure (OECD 2009, 84).

Definition in private sector R&D. In examining the emergence of new technologies, identifying the value of private sector spending is crucial as this represents the largest total investment in biotechnology-related R&D.[5] Morgan Jones and Grant (2011) show the total UK public sector spending on health related R&D was £1.5bn in 2008/2009, while biomedical charities spent £1.1bn, and private industry invested £8.9bn. Consequently the question of whether it is possible to identify how different technological options are financially supported therefore primarily becomes a question of whether it is possible to track the R&D expenditures of a relatively few large pharmaceutical firms (the top 15 US/EU firms spent US$50 billion in R&D in 2009 – Rafols et al. 2012). Yet the private sector's systems for accounting for spending on new technologies is opaque making it very difficult to identify levels of investment in specific areas.

For example one interviewee suggested that pharmaceutical firms tend to begin accounting for technologies once they reach clinical trials rather than before and when asked to estimate their spending by types of market to estimate their spending on orphan drugs he recalled:

> they told us it would be difficult to disentangle that information internally, they wouldn't recognise what was orphan and non-orphan … they didn't know how to apportion it internally.

These results do not reflect confidentiality (although this may be relevant too in some cases) but the interviewee maintained, this was because of the motivation for collecting the data in the first place as he explained:

> if they [big pharma firm X] have a research facility at [town Y] all they need to know is how much it costs to keep going. Sometimes those facilities coincide with projects in individual groups of

> technology or individual technologies if you are lucky but most of the time they don't. They never actually have the information in the form you need it for the type of [report] you are talking about.

In this context the thorough identification of funding for particular technologies, this becomes very difficult. To the firm, a proposal to measure investment in a type of technology is entirely novel as this is not a classification system they use, and the analyst collecting such data has to build their estimate from scratch as well. Consequently the firm and the researcher are both confronted with novelty (see also Section 4.1.2).

Definition in public sector R&D. Within the public sector, there is considerable scope for plurality among the different funders who to categorise investments in their own biomedical R&D. The definitions used by one organisation's funding regime may differ considerably from that of another in the same country or within the organisation over time. Leaving aside the issue of large differences in the response rates in different countries who were invited to contribute to research studies (Section 4.2.2 revisits issues such as incentives, and problems of working across multiple jurisdictions). However even where there was support, classifying information even into broad sectors of application was difficult: 'She doesn't use our categories, we don't use hers' one interviewee concluded looking back at material gathered for their empirical study. Perhaps it is inevitable that as a result the interviewee recalled 'Most respondents leave empty the specifications of percentage allocated to different funding categories by application areas covered and activities.' However, the interviewee suggested that funders could distinguish qualitatively the goals (for example firm creation) and activities their funding aimed to support, for example if the funding was meant for applied research or basic research.

4.1.2. *Substance: emergence*

An initial starting point in analysis of a new technology relates to the identification of the new technology. In this case there are two key issues identified previously, in line with Power (2004, 2005): novelty and performativity. As a new technology emerges, a framework must exist for it to be worthy of identification as being new. The challenge of capturing novelty is fundamentally difficult for several reasons. First, there is likely to be a lag between work commencing on a technology and the new classification system being developed. Second, the new phenomena may initially be classified as a subset of some existing classification at least until it is given its own grouping. For example, to find financing via corporate alliances for the emerging technology of RNA inference using Deloitte's Recombinant Capital database, in recent years it has been necessary to look under 'antisense technology', a prior technology that has a similar (but distinct) operational principle. Third, novel activities may be actively hidden and unreported even to those managing the organisations that support them especially where these are utilising funding that has not been approved (Augsdorfer 1996).

Novelty is an issue for public sector organisations too. Consider the most accessible dataset on funding of biomedical research is provided by the US National Institutes of Health (NIH). The NIH distributes over US$30bn annually in research, mainly to hospitals, universities and its own laboratories. It is possible to explore the allocation of funding since 2008 using the NIH Research Portfolio Online Reporting Tools (RePORT). However there are some limitations to the system and only data on pre-determined categories are provided. The mix of diseases, technologies and disciplines that form the current set of categories is not comprehensive and is uneven in coverage (see Figures 2 and 3). The types of technologies that can be searched are shown in Figure 2. For example while it is possible to find money spent on genetic testing, it is not possible to search for diagnostics overall. There is no category for pharmaceuticals or sub-types of drugs,

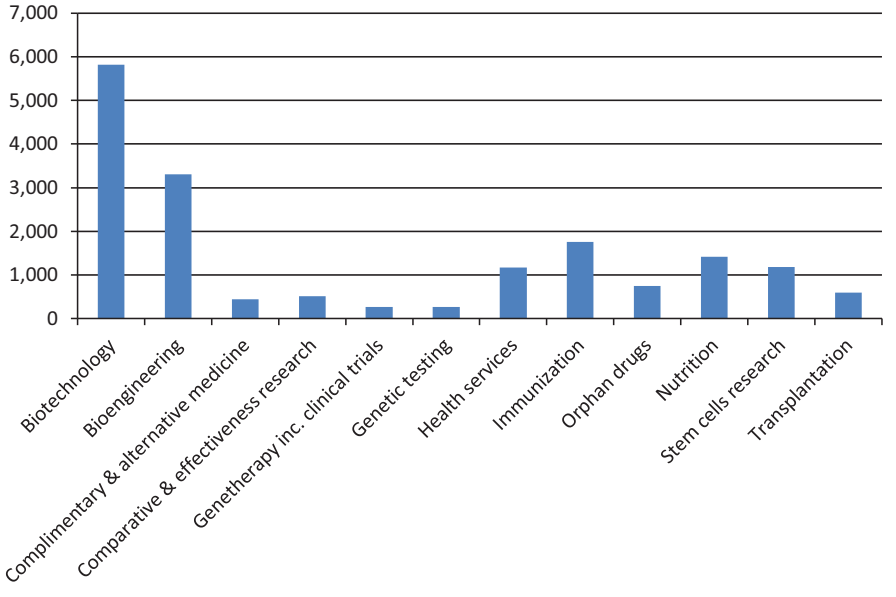

Figure 2. NIH funding (US$m) in 2011 in selected non-disease specific fields.

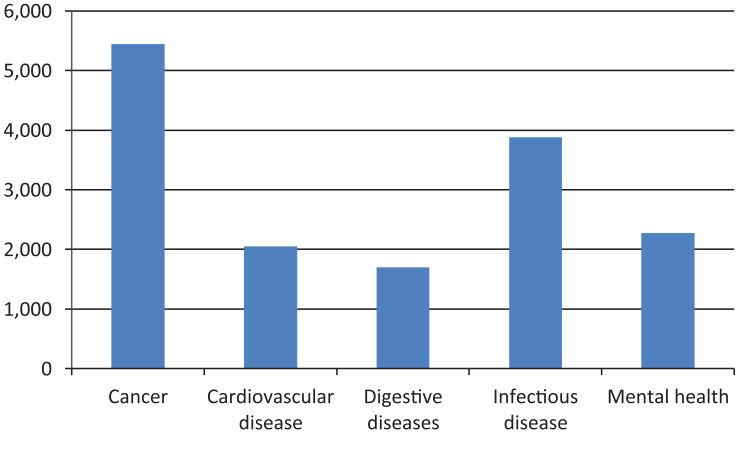

Figure 3. NIH funding (US$m) in 2011 NIH in selected disease areas.

such as monoclonal antibodies, although orphan drugs have a category. There are overlapping categories for regenerative medicine and stem cells, immunization and vaccines, but none for medical devices.

The analytical typology reflected the NIH in RePORT reflects 'categories … that were, over time, requested by Congress and other Federal agencies for reporting to Congress and the public' (NIH FAQs). The categories available therefore reflect those fields that are of political interest, in broad terms, and not an analytical technological typology. This reflects the inherent performativity associated with identification of new technologies; the categories used reflect the political priorities

of those funding research, and these interests are then reflected in what is funded. Because Figure 3 shows a small number of relatively non-overlapping disease categories, chosen for illustration only (but many sub-classes especially of cancer are available) as well as levels of funding support for areas that cut across diseases including technologies and research approaches (and will therefore overlap substantially, for example gene therapy would be entirely contained within biotechnology. The use of categories that reflect political priorities means that other, new technologies will not be reflected in the given system until they are identified as a new phenomenon worthy of categorisation by politicians or policymakers – only at which point will the technology be officially observed.

4.2. *Quantification*

While the previous section highlighted the challenges with identifying clear-cut categories, even if such categories existed there would be numerous problems with the actual quantification of R&D investments. This section highlights three thematic problems around quantification – processes of counting, gathering, and degradation of over time – around which we frame the issues to quantification identified in Section 2.

4.2.1. *Processes of counting*

For the quantification of biotechnology R&D funding data, it would be overly simplistic to suggest that an obvious, objective means of counting exists. Consequently the issue of methodology is non-trivial over many dimensions. There are questions of quanta – units of analysis – and the norms of data counting such as at which threshold a phenomenon is deemed to be observed or worth counting. Also there remains considerable room for judgement by researchers about what to include and which category to put observations in. For instance, simply identifying the number of firms or institutions conducting research in a given field can be relatively difficult. As discussed above, 'biotechnology firms' represent boundary objects that can include a range of firms conducting any number of biotechnology-related activities. Because there is no way of knowing which firms are conducting these activities without asking, there is an instantaneous problem simply in identifying where to start. Therefore, simply counting biotechnology firms becomes more difficult, as researchers must make a subjective judgement about the point at which a firm may have sufficient 'biotechnology' activities to be considered as a biotechnology firm.

The private sector is not unique in being difficult to identify actors. The same problem exists in the public sector, where the sheer number of funders – even in a relatively small country like the UK – makes finding a method for merely identifying those funding research problematic. For example, Morgan Jones and Grant (2011) undertook a study for the UK Department of Health to identify funders supporting the relatively small field of complex trauma, having initially identified the main funders of UK public-sector R&D. This was achieved by first undertaking interviews with scientists in the field and then a bibliometric analysis, using key words, to identify publications in the field to track authors of relevant research and finally find out who had funded this work. The need to find funding support mechanisms within government departments, at national and regional levels, as well as charitable and large and small private firms makes this time consuming. In other instances firms or public sector organisations may be required to respond to surveys to report activities which may collated in national statistics (Patel 2012).

Table 2 shows that when the OECD (2009) attempted to gather data on public sector investments in biotechnology from countries, few members responded and in those that did the method was different in almost all cases as these differed in sample frame, whether responses are mandatory or non-mandatory and by extent of coverage.

Table 2. List of countries providing data on aggregate R&D spending (all global data available)

Country	Year	Biotech definition	Sample frame used	Extent of coverage	Mandatory?	Who conducted survey
Canada	2005/6	OECD	Census	Federal govt departments and agencies performing science and technology (S&T) or with budgetary allocation to fund S&T	Yes	Govt
Czech Republic	2007	OECD	R&D survey	All	Yes	Govt
Korea	2006	All	R&D survey	All	No	NIFU STEP
Norway	2005	OECD	Census	Total higher education sector and the Norwegian Institute sector	No	Govt
Poland	2005	OECD	Secondary sources	Selected for S&T	Yes	Govt
Spain	2006	OECD	Census	Only government institutions with S&T activities or budgets	Yes	Govt
Slovenia	2005	OECD	Secondary sources	Selected for S&T	Yes	Govt

Source: OECD (2009).

Furthermore studies sponsored at a national level may focus on national institutions, but a national innovation system may benefit from international inputs at multiple points, such as research grants and multinational corporate R&D (Enzing et al. 2008). However in practice studies have not tracked international influences as one interviewee noted:

> in terms of research outcomes it is actually impossible to take into account international influence … we all recognise that it is there but I am not aware of anyone taking [it] into account … other than assuming it away [i.e.] just looking at the US and saying let's assume the US equals the world because it is a dominant player in medical research and a very large economy by itself.

These factors combine to make it difficult to establish a clearly auditable trail linking money actually spent to figures generated for biotechnology R&D. Different organisations may have their own audit practices whereby the requirement to produce data year after year generates substantial expertise and will improve the internal consistency of data – for instance as the US NIH RePOST system is doing. However the range of methods and systems for capturing these practices makes it very difficult to clearly identify ways of making the data from different systems sufficiently compatible to be able to draw meaningful conclusions about new technologies across the jurisdictions of different funders.

4.2.2. *Data gathering*

For all the problems of counting (even sources of) biomedical R&D identified above, there are also several more practical problems with the collection of substantial meaningful data. Issues of

incentives, jurisdiction, confidentiality and cost also impact the ability of researchers to collect useful data about R&D spending.

The ultimate aim of any analysis of R&D spending is to draw a meaningful conclusion about the level of investment in a new technology, but aggregating data from different sources in a form in which it may be successfully analysed is fundamentally difficult. Beyond the issues discussed above regarding the establishment of categories and methodologies, there are fundamental issues of incentives and jurisdiction, which limit the ability to aggregate data without bias. The issue of incentives is key to data collection – the collection of data is predicated on any particular agent's willingness to give correct information. For private sector organisations, even if they could classify their own R&D spending on emerging technologies (which, as discussed above they often cannot), incentives to respond to researchers, or even to produce information for other purposes that researchers can draw on, will vary.

An example of independently produced information that researchers can use would be press releases or on company websites, but these are subject to reporting biases as a result of varying incentives depending on ownership types, age, etc. Further, while some organisations (public sector or private) may be compelled by legal or political reasons to disclose information, others may be less willing, contributing to an overall response bias that is intractable, as the information not provided is difficult or impossible to find elsewhere.

Coupled with issues of response bias and incentives is the broad challenge of data gathering owing to jurisdiction and confidentiality (already mentioned above) and cost. There are, as mentioned in the previous section, considerable issues with substance categorisation and methods among different countries. Furthermore, accessing data and participants on an international scale often necessitates work with foreign partners who have access to data. The scope of data collected is determined by the definition of the field as established by the analyst, but this may not be reflected in the data collection systems within different countries. Consequently data has to be discussed and collected through an interactive process with funders – who may or may not devote their time to helping with such enquiries. Consequently, without the ability to shape the collection systems and rules for quantification, different figures (such as those in Table 1) are produced and conflated, making the completeness of 'aggregate' figures potentially questionable and at times outright disputed (for example, see Arundel, van Beuzekom, and Gillespie 2007).

Compounding this problem is the issue of confidential data – for the long-term study of the emergence of biotechnology there are commercial information providers such as Ernst and Young, who have longstanding datasets that have tracked trends in biotechnology, with an annual Beyond Borders report and contributions to annual reviews by journals such as *Nature Biotechnology*. Ernst and Young data is gathered and maintained by what appears to be a large team of analysts, but is proprietary. While these data may be available to explore for commissioned work such as business intelligence provision, this comes at a cost.

Obviously, it is important to consider the crucial logistical issue of cost. As valuable as data on R&D spending may be for a particular area, the costs of collecting or accessing this data must not outweigh its value. Unfortunately our interviewees highlighted the extensive time costs of pursuing their own data collection. Morgan Jones and Grant's (2011) study of the complex trauma examined a small field (receiving perhaps £15 m per year), but the process of undertaking interviews with key researchers, bibliometric analysis, funding mechanisms and funding sources meant that even though funding in only a single financial year was recorded, the exercise out took several weeks of research. The extension of the project to other areas or countries increases the complexity many times over; for the report in Enzing et al. (2007) gathering data for public sector

spending on biotechnology took months per country, even for small European countries, and the reliability of the data was uncertain.

4.2.3. *Degradation of data over time*

One final issue is that of the dynamics of spending data over time, which relates both the transformation and loss of data. Given scholars' interest in mapping the long-term trajectories of new technologies, data extending over a period of time is desirable, thus changes in the nature of technologies over time are of importance.

When research begins on a new area, its ultimate purpose may not be clear: as our interviewees commented, spending by pharmaceutical companies on a particular technology is difficult to track because 'in preclinical work [pharmaceutical firms'] investment is not technology specific even if you have access [to interviewees familiar with the technology]'. Further, its utility may in fact undergo a transformation: 'you might find something different than you set out to find … look at anti-retroviral drugs for HIV/AIDS, they were failed cancer drug'. Consequently expenditures may be recorded differently depending on when an analyst had asked what money was spent on the first generation of the drug.

Another factor that occurs over time is loss of data, for example, because of recall bias. As time passes, the memory of projects and their origins can shift, as one interviewee commented:

> … there is a difference in the ex-ante and the ex-post view. Ex-post you say this is what we produced and everything we spent on R&D produced this … . Ex-ante you may think you were spending on a whole range of stuff most of which never made it to the market.

These difficulties mean that it is important to capture research on emerging technologies as it happens, as *ex-post* investigations run the risk of mis-capturing the real expenditures on a range of new technologies over the extended period of time.

4.3. *Summary*

This section has described how almost all of the issues anticipated in Section 2 (with the exception of quanta) have been observed in a review of national and international surveys related to tracking investments in biotechnology or biomedical sciences. In the following sections we explore whether proxy measures can provide a way to avoid these difficulties.

5. Proxy measures and tech mining as alternate means of tracking emerging technologies

Having previously focused on the challenges and benefits of tracking R&D spending, now we similarly explore the properties of proxy measures as means of tracking the evolution of emerging technologies and related capabilities and research areas. It is important to contrast the definitional differences between R&D spending and proxy measures; while R&D spending examines inputs to the innovation process, proxy measures tend to track outputs in various forms. There are a wide range of outputs of innovative activities that may be used as proxies for tracking innovative activities – see Patel (2012) for a review. However a suitable proxy for R&D investment should be close to the act of investment itself and be able to cover the scope of activities that investments may support (*ibid.*). Consequently, in the field of emerging biotechnologies, publications and patents address this area well as these are early outputs in the innovation process: they cover products or processes (both incremental and major), and they can be collated for a wide range of analytical

levels (country, industry, technical field, organisation). Patents and publications can reveal data on small and large organisations and data is collected internationally and independently (*ibid.*).

The past 20 years have seen considerable advances in the use of publications and patents as tools for academic inquiry into science. Bibliometric techniques based on publications, while originally used as means of measuring scientific excellence, have also been used increasingly to proxy innovative activities and examine the emergence and evolution of technologies (Veugelers 2005). Given that publications are more likely to come from academic sources (Murray 2002), publications and citation patterns can be particularly useful for measuring dynamics of scientific research (Leydesdorff 1987; Boyack, Klavans, and Borner 2005; Porter and Youtie 2009). With this said, some industries – particularly pharmaceuticals – rely on publications for external validation of technology (Rafols et al. 2012). Similarly, patents identify in detail the state of a particular technology, at a particular point in time (Jaffe and Trajtenberg 2002). Because of their requirement to demonstrate novelty, patents have been widely used as a reflection of organisational R&D capabilities (Rosenberg 1982; Pavitt 1985; Archibuigi 1992; Jaffe et al. 1993; Jaffe and Trajtenberg 2002). In this section we will examine patents and publications utility for tracing investment in emerging biotechnologies using the framework discussed in Section 2. We consider these proxy measures in terms of 'traditional' bibliometric methods and also discuss how tech mining techniques may help to mitigate weaknesses of these techniques. We suggest that while proxy measures have potential to overcome some of the weaknesses inherent in funding data, they come with their own biases and logistical issues that must be considered. We then propose that tech mining may be a useful means to address some of these problems.

5.1. *Substance*

5.1.1. *Substance: definitions*

Given the critiques of R&D funding listed above, both patents and publications appear highly appealing as means for tracking emergence of technologies. Both utilise clearly defined multi-level indexing categorisation systems whose high-granularly and uniformity make them useful for analysis. Patents are granted and indexed using standardised codes to identify technologies and the relative position of a technology by professional patent examiners with expertise in the area in question. Publications are organised around journal articles, which focus on specific focused areas and are edited, reviewed and indexed by experts.

Consequently these classification systems may help to address the issues of plurality, indistinctness and boundary objects by operating within specifically identified categories, for instance patent codes or journals. These categories may serve as units of analysis themselves, for instance patent codes (Hall, Jaffe, and Trajtenberg 2001; Strumsky et al. 2012) and journals (see Glanzel and Moed 2002). The parallel issues of indistinctness of new technologies and boundary objects around unclearly defined topics are addressed by two key factors prevalent in both patents and publications. In addition to the detailed categories for the classification of new patents or publications (see uses of these for example in Strumsky et al. 2012; Rafols et al. 2012; Leydesdorff et al. 2012), both patent and publications acquire citations, which document how a unit of output relates to previous work. In this way new technologies may be observed from the bottom-up as new technologies may be traced from the previous advances they cite. In theory such search methods can generate more sophisticated and granular results than R&D funding data. For example, while the term 'biotechnology' is used in a loose way that serves as a boundary object and leaves the analyst reliant on funder's definitions of technology, when using proxies it is possible to identify

all papers or patents citing a specific biotechnology technique or classified in a particular field by indexers that are independent of the funder.

However while the premise of categorisation and citations for publications and patents may appear simple, these categories include a range of opaque, socially mediated factors that are not immediately observable and introduce important biases into our ability to track emerging technologies. For instance, Benner and Waldfogel (2008) find that using only the first listed technology code on patents to characterise a patent may mis-state the proximity of technologies, and publications databases such as ISI rely upon human judgement about content to categorise journals (Boyack, Klavans, and Borner 2005). Meanwhile the use of citations also has logistical issues: patent citations come from both the applicant and the examiner, obscuring efforts to examine the impact or awareness of prior innovations (Alcacer and Gittelman 2006) and for publications mis-matching of citations is rife (estimated by Moed (2002) to include up to 30% of citations in some cases), introducing a slightly different element of *indistinctness* to the issues discussed for R&D statistics. In addition, it has been argued that citations themselves may be interpreted as a boundary object, making the rigorous unit of citations less concrete than they might initially appear.

5.1.2. *Substance: emerging technologies*

Access to citations also means that the identification of novelty of new technologies is considerably easier. The bottom-up nature of patents and publication citations means that new technologies may be tracked from their origins, via the originating papers or patents. Importantly, mechanisms exist for new categories to come into being: patent offices review and add to coding systems over time, while new journals may be created to reflect academic advances, and clear procedures exist for journals to be added to the main databases.[6]

Again however, there are logistical issues that cloud the effectiveness of these techniques; delays in the review and publication of results can introduce time lags for the measurement of new technologies. Similarly there are lags for the review and introduction of new journals. Larsen and von Ins (2010) suggests a decline in the total amount of scientific activity captured in the major journal indexes. This may take the form both of new journals, as well as other activities in the form of working papers, conference papers or other communications outside the realm of indices (Veugelers 2005). From the perspective of tracking new technologies, this means that outputs only appear once they have been sufficiently advanced to become outputs; the lag between input measures such as funding and these outputs are therefore potentially troublesome.

Performativity is an issue as well for publications, which are particularly susceptible to game-playing activities such as self-citation, coercive citation and pressures in journal selection. Research may be self-declared to fit into categories that are advantageous for authors to be classified in (see Calvert 2006).

5.2. *Quantification*

5.2.1. *Quantification: method*

The process of quantifying investment in new technologies was discussed earlier as being fraught with difficulties for the study of biotechnology R&D. This process is, superficially, considerably easier for proxy measures. First, the quanta or unit of analysis, is clear: a publication or patent is the basis of any counting. The method for collection and analysis of data is similarly straightforward, as databases of patents and publications are available easily and cheaply, and

maintained by dedicated professionals over long periods of time for primary purposes other than monitoring technological investments (which provides some neutrality with regards incentives). It is possible to download thousands of citations for immediate advanced analysis, for which there is extensive documentation of research methodology in the literature. The norms for inclusion are clearly recognised standards: for patents the thresholds of novelty, inventiveness and disclosure, are clearly defined and codified in legislation and case law. For publications the established peer review system determines which is published, and clear guidelines exist regarding which journals are included in major databases (see ThomsonReuters 2012). Once a researcher is conducting analysis on either patents or publications, many questions of judgement may have been made by professional examiners or indexers. Finally, the use of patents and publications to track technologies feeds upon inherently auditable mechanisms required to make both systems robust to legal or academic scrutiny.

Again, however, there are theoretical and practical issues posed by these proxy measures. In terms of threshold and quanta, not all innovations are patentable, and of those that are patentable not all reach the threshold to be granted (Hall, Jaffe, and Trajtenberg 2001). Even those meeting patentability criteria may not be granted for economic reasons. Similarly the political processes inherent in the peer review system means thresholds for acceptance to journals are uneven (Hojat, Gonnella, and Caelleigh 2003). The judgement by the analyst is replaced by the (still subjective) judgements of patent examiners and indexers involved in classification of journal articles, meaning researchers have less control over the terms used in quantifying these innovative activities. In addition to all of these, the methods required to track emerging technologies in patent or publication databases are made more difficult by challenges in cleaning, matching and preparing data that may not have originally been intended for this type of analysis.

5.2.2. *Quantification: data gathering*

The mechanics of drawing conclusions from proxy measures represents both benefits and drawbacks. The advantages are significant: the cost of accessing patent and publication data is relatively low, both for the subscription to the databases (at least assuming academic analysts) and for software required for analysis. Incentives to participate are also addressed by the design of the patent/publication systems – the observation of innovative activity using patents or publications is a secondary consequence of individuals and firms' desire to protect and/or publicise their work. These rewards encourage the disclosure of information that would otherwise be kept secret. This means that confidentiality is less of an issue for proxies than for R&D figures, as one key element of patenting or publication is the disclosure of information in return for rewards. Of course the public nature of these disclosures does mean that some innovations will not be captured, either because they cannot be patented, because of commercial sensitivities about appropriation, or because of security concerns (i.e. 'black patents'); however for these it is not clear that data on R&D investment would be any more illuminating.

There are other issues associated with collecting of data. Incentives to publish – particularly for private firms – are not necessarily driven by a straightforward desire to engage with academic debate (Sismondo 2009). Consequently the publication activities of firms may at various points overstate (Hicks 1995) or understate (Sismondo 2007) the actual level of research taking place. Perhaps a greater concern for both patents and publications comes from the issue of international data gathering, which presents a different facet of the issue of jurisdiction. Propensity to patent is not consistent across countries (especially developing countries) or industries, e.g. software, services (Archibuigi and Pianta 1996). Moreover, patents represent the result of a process externally

mediated by patent examiners, and reflect national contexts, which result in different citation patterns (Meyer 2000; Alcacer, Gittelman, and Sampat 2009), while the costs in obtaining granted patents may mean smaller organisations or academic organisations are less active (Hopkins et al. 2006). For similar reasons we can expect those from developing countries will be less able to afford large patent portfolios. Consequently, conclusions about cross-national activities require caution as the rules and citation patterns nationally may be fundamentally different. A similar issue exists for publications, where non English-speaking countries may have non-English publications that are not captured or are undercounted in mainstream databases and may reflect different citation patterns, which may then impact on other research (see Van Leeuwen et al. 2001). In addition to these issues, there are more basic logistical issues posed by work with proxies; while costs of access are relatively low, there are challenges posed by limits to downloads, restrictions on database access (especially to non-academic analysts) and other matters that can significantly impede a researcher's progress.

5.2.3. *Quantification: decay of data*

The issue of the loss of data with memory is effectively avoided with the use of proxy measures such as patents and publications in the digital age because patent and publication data now exist in accessible forms covering long periods of time, it is possible to track the long-run evolution of capabilities over time (see Jaffe et al. 1993; Gittelman and Kogut 2003; Hess and Rothaermel 2011). The transformation of technologies may, in these circumstances, be observed and studied rather than be the source of uncertainty about the counting of bits of data. With this said, transformation may only be possible to observe in cases where the keywords and terms for technologies remain constant over time; in these cases there is potential for data to be lost if traditional bibliometric techniques (using titles, authors, keywords, etc.) are used.

5.3. *The prospect of tech mining for improving publication and patent analysis*

Our previous discussion has mostly focused on the uses of proxy measures and the implicit use of traditional techniques to analyse them. However it is important to emphasise the potential held by new tech mining techniques based on analysis of co-occurrence of proximate terms within text to improve the ability to generate insights from these proxy measures. Tech mining approaches show considerable potential to build on advantageous characteristics of proxy data – for instance large and low-cost datasets that are indexed independently of funders, but to mitigate some issues identified above. For instance, tech mining techniques may be used to identify larger and more precise datasets related to novel technologies, perhaps from an earlier stage than might otherwise be observed using publications indexed in the traditional manner, and data mining techniques have potential to aid in the observation of transformation of technologies over time (Yang et al. 2008). Use of tech mining to find semantically related research allows for the mapping of the emergence of technology in a more sophisticated way than had been done before (see Pudovkin and Garfield 2002; Morillo, Bordons, and Gomèz 2003).

5.4. *Summary*

Against the criteria that we identified in Section 2, it is clear that many of the problems that made use of biotechnology R&D figures troublesome are at least partially resolved by the use of proxies for tracking technologies, although these measures are far from perfect. Patents and publications, despite their limitations, are considerably more adaptable and suited to objective quantification

than are politically charged figures for funding, and all the attendant problems associated with them. With their weaknesses these proxies are ultimately complementary, and can be used together to offset weaknesses and to provide the possibility of multiple partial indicators which, when giving convergent results, provide a stronger signal for policy and strategy deliberations (Martin 1996).

6. Combined funding measures and proxies in action – the emergence of genomics as a drug discovery technology

To support the claims of the previous section, here we present an example of the different conclusions that may be drawn from input–output and proxy measures of biomedical R&D. Around the time of the Human Genome Project (1990–2003) and for some years afterwards, high expectations surrounded techniques that allowed the sequencing of genes, and analysis of their structure and roles in relation to the biological processes linked with disease. These genomic technologies were seen as crucial new tools for drug discovery efforts across the pharmaceutical industry that would allow identification of the root causes of pathologies and facilitate the design of more precise molecular interventions (Martin et al. 2009).

During the 1990s scores of dedicated biotechnology firms were founded specifically to exploit genomics, first using technology platforms to identify potential disease mechanisms, and later through development of their own drugs based on genomic insights (Rothman and Kraft 2006).

In 1999 as firms and public organisations raced to sequence the human genome and technology financing was in a boom phase, the New York Times interviewed William Haseltine, the CEO of one leading genomics firm, Human Genome Sciences, who provocatively stated: 'Any company that wants to be in the business of using genes, proteins or antibodies as drugs has a very high probability of running afoul of our patents …. From a commercial point of view, they are severely constrained – and far more than they realize'.[7]

In light of the apparent progress of the genomics firms, the stakes seemed extremely high as a new sub-sector of the pharmaceutical industry emerged. By 2000, the top 5 genomics firms were valued at US$45bn on US stock markets (collectively equivalent to the size of a top pharmaceutical firm). It was (and remains) relatively straightforward to assess the progress of these would-be challengers to the incumbent pharmaceutical firms, given the main focus of these firms is exploiting genomics and the resulting products and revenues stem from investments in this field. Such measures as R&D investment, progress of products through the drug development pipeline (including historical views) are shown in Table 3 for leading genomics firms, as collected from Securities and Exchange Commission filings.

By contrast when looking for comparable data from pharmaceutical firms, we find only firm-level data on R&D spending as a whole. It is not possible from an external vantage, or even perhaps internally, to use the same measures to tell which of the many products under development are the result of investments in genomics (which can support the discovery of different modes of therapy) or how much of total R&D spending is devoted to this activity. However annual reports, press releases and corporate alliances all gave signals large pharmaceutical firms were active in the field. Commentators suggested estimates of annual spending of US$100–300 m based on discussions with firms (Gassmann, Reepmeyer, and von Zedtwitz 2004), although it should now be clear from the discussion in Section 2 these estimates will be subject to many uncertainties.

It is here that proxy measures provide an opportunity to draw direct comparisons with some precision and to ask the question, to what extent are pharmaceutical firms keeping up with new entrants in the field of genomics?

Table 3. Measures of progress in leading genomics firms

Company name	Revenue (US$m) 2004	Revenue (US$m) 2008	R&D (US$m) 2004	R&D (US$m) 2008	No. of molecules in pipeline (2010)	Lead compound stage (2010)
Human Genome Sciences	3.8	48.4	219.6	242.7	8	Market
Rigel	4.7	109.7	48.5	8.0	8	Phase 2
Sangamo Biosciences	1.3	16.2	11.1	31.2	1	Phase 2
Maxygen	16.3	100.7	53.3	46.3	3	Phase 2
Ariad	7.4	8.75	27.7	50.8	3	Phase 3
Incyte	14.2	3.9	88.3	87.6	12	Phase 3
Lexicon Pharmaceuticals	61.7	32.3	90.6	108.6	5	Phase 2
Exelixis	52.9	117.9	137.8	257.3	14	Phase 3
Arena	13.7	9.8	57.7	204.4	5	At FDA approval

Source: SEC filings.

6.1. *Patents as proxies*

Patents claiming DNA sequences can be selected directly from patent databases such as patent lens or specialist indexers such as Thomson Scientific, without the need to search specific patent classes (see method for collection in Hopkins et al. 2006). These are used here as an appropriate a proxy for investment in genomics as these are gained only for novel inventions (a sign of capability) and novel DNA sequences are only accessible using genetic techniques (or else indirectly deduced using slower but closely related protein sequencing approaches) and so these are reasonably specific proxies. The costs of writing and obtaining a US patent in biotechnology are relatively inexpensive, perhaps US$10,000 compared with Ł55,000 for a granted patent at the European Patent Office (Hopkins et al. 2007) thus few firms should be excluded from visibility by using this measure. Furthermore all firms engaged in genomics had, at the time, a strong incentive to patent as access to future drug targets was felt to be at stake (although Hopkins et al. (2007) note the intellectual property regime around this particular field has weakened since 2001 and such changes are an important contextual element to be considered in selection of proxies). Patents claiming DNA sequences are selected as suitable proxy-measure in this instance because these are found to be routinely generated as a product of research by many firms, in an industry with a high propensity to patent. Furthermore, US patent grants are selected as the USA is the world's largest market for therapeutics thus firms looking to exploit genomics for therapeutic purposes could be expected to focus attention here – indeed most genomics firms were founded in the USA (Martin et al. 2009).

Table 4 distinguishes between dedicated genomics firms and other (pre-genomics) biotech firms, as well as incumbent pharmaceutical firms (noting any prominent acquisitions of genomic-capable firms whose patent estate is counted towards the acquirer's total). The findings are revealing, showing that (despite Haseltine's claims) there are more pharmaceutical firms in the top 10 holders of DNA sequence patents than genomics firms, and indeed these firms and others such as biotech firms e.g. Amgen had been active in the field before the genomics firms were established.

Table 4. Top 10 corporate holders of DNA patents granted in the USA

	US Patents granted (by application period)				
	All years	1980–1990	1991–1995	1996–2000	2001–2003
Incyte Corp (genomics firm) USA	572	3	22	529	18
AMGEN INC (biotech firm) USA	290	28	85	149	28
Human Genome Sciences (genomics firm) USA	289	0	83	167	39
Millennium Pharm Inc. (genomics firm) USA	260	0	19	196	45
GlaxoSmithKline (pharmaceutical firm) UK	228	2	13	200	13
Isis Pharm Inc. (biotech firm) USA	227	0	6	176	45
Roche (inc. Genentech) (pharmaceutical/diagnostics firm) Switzerland	222	21	66	111	24
Applera Corp. (includes Celera) (instrumentation/genomics) USA	162	0	3	37	122
Wyeth (pharmaceutical firm) USA	153	18	43	83	9
Novartis (inc. Chiron) (pharmaceutical firm) Switzerland	142	16	57	64	5

Source: Hopkins et al. (2006), 24.

6.2. *Scientific publications as a proxy measure*

Publications provide an alternative but complementary window on large pharmaceutical firms' activities in genomics. Large pharmaceutical firms typically publish *ca* 1000 papers annually each thus providing an opportunity for analysis of broad-based trends (Rafols et al. 2012). Bibliometric approaches that use keyword searches can provide high-granularity to distinguish firms as shown in the example below. The method employed in the Tables 5 and 6 (previously in Hopkins et al. (2007) – a fuller description is given there) uses search terms that describe genomics-related techniques that are derived from review of contemporary scientific papers and counting only research papers (i.e. excluding review articles, editorials, etc.) to ensure that search results indicate technical activity, rather than simply reflection pieces. Tables 5 and 6 show publications for each firm by year, shaded in grey where the firm is active (1+ publications per year) darker shades reflecting higher activity (the darkest being >20 publications per year). Unshaded years indicate external activity (the field existed but the pharmaceutical firm did not publish), with a '0' indicating discontinuous publishing, while an 'X' indicates the field had not emerged at that time and so publication was not yet possible.

While patents are often necessary to commercially exploit scientific developments in drug discovery, publications may not be, and so propensity to publish scientific papers between firms varies considerably. For example data from Rafols et al. (2012) shows AstraZeneca and GSK, the two largest UK pharmaceutical firms, and both in the global top 5 by sales, spent an average of £2.7bn and £3.1bn respectively on R&D over 2004–2007, but although GSK spent only 15% more it published 33% more papers (AZ published 3282 vs GSK's 4355 papers). Despite apparent differences in propensity to publish in general, Tables 5 and 6 show that GSK published 1487 papers within the defined keyword set, compared to AstraZeneca's 331. This is consistent with the patenting position, where AstraZeneca despite being in the top 10 R&D spenders in

Table 5. Accumulation of research interests related to genomics at GSK 1991–2003

Subject	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003
Recombinant DNA	3	1	4	2	2	1	1	0	11	8	3	2	1
Sequencing DNA	7	12	9	15	13	25	20	21	14	41	12	13	15
Gene cloning	18	18	16	15	17	18	14	11	16	25	9	9	5
Protein sequence	20	21	20	27	27	42	29	38	28	39	11	25	20
Protein expression	1	3	3	2	2	3	6	10	6	10	2	11	9
Gene expression	8	6	6	11	17	9	25	30	26	28	21	26	21
Sequence homologies	6	1	4	2	0	4	5	5	7	6	0	7	4
Transgenic animals		1	1	2	1	5	5	6	5	7	4	2	2
DNA database	3	5	5	5	6	8	7	12	6	7	5	3	1
Bioinformatics				1	0	2	3	4	4	11	12	10	9
Gene function					1	1	3	1	5	2	7	3	1
Genome mapping					1	0	1	1	2	4	2	2	1
Gene knockout	X				1	0	1	2	1	2	1	1	0
Genotyping						2	2	0	2	5	1	2	4
Population genetics						1	1	1	1	5	0	5	0
Pharmacogenetics						1	0	1	1	7	2	6	2
Microarray							1	0	1	7	4	7	7
Proteomics	X	X	X	X	X	X		1	2	9	6	6	6
SNP analysis	X	X	X						1	4	2	4	2
RNAi	X	X	X	X	X	X	X					1	0

Table 6. Accumulation of research interests related genomics at AstraZeneca 1991–2003

Subject	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003
Recombinant DNA			1	0	0	0	0	0	1	0	0	1	0
Sequencing DNA	7	10	3	2	3	4	2	5	5	2	1	2	2
Gene cloning	5	8	3	7	1	5	0	8	1	4	2	2	1
Protein sequence	3	9	4	4	6	4	3	5	4	5	5	5	8
Protein expression	0	2	0	0	1	0	0	0	2	0	2	2	8
Gene expression	2	7	1	3	3	4	4	8	6	7	2	3	4
Sequence homologies	1	1	0	2	0	0	1	2	0	1	1	0	0
Transgenic animals	0	2	1	4	3	1	1	3	0	2	2	2	0
DNA database	1	2	0	1	0	1	0	1	0	0	0	0	1
Bioinformatics										1	2	2	4
Gene function	2	0	0	0	1	0	1	0	0	1	0	1	0
Genome mapping	1	2	0	0	0	0	0	0	1	0	0	1	0
Gene knockout	X								1	0	0	0	0
Genotyping	1	0	0	0	0	0	2	2	0	1	0	2	0
Population genetics	0	0	0	0	1	1	0	0	0	1	0	1	1
Pharmacogenetics										1	2	2	0
Microarray								1	0	2	1	2	3
Proteomics	X	X	X	X	X	X				2	3	3	4
SNP analysis	X	X	X										
RNAi	X	X	X	X	X	X	X						

pharmaceuticals (see Rafols et al. 2012) is outside the top 10 holders of DNA patents. Therefore publication levels suggest that GSK was undertaking much higher levels of research in fields related to genomics than AZ. This position is further supported by the finding (reported in Martin et al. 2009) that AZ had only half the commercial alliances in the field compared to GSK, based

on firm-firm deals announced in the press over the decade 1993–2003.

The contrast between the level of detail presented in input (funding) vs output (proxies) is considerable. We are able to gain much richer and more detailed information about the actual research activities of the firms in question using the proxy figures in contrast to the figures for funding which are only available as estimates by internal company staff.

7. Discussion

This paper has presented and tested a taxonomic framework for examining the strengths and weaknesses of several methods for tracking R&D activity related to the emergence of technologies. Using this framework we explore primary and secondary accounts of data collection efforts revealing at least 16 distinct issues that present problems when attempting to follow which organisations are investing in different forms of technology and the extent to which they are participating in such activites. This is particularly difficult when we consider that technologies may develop on a global stage and so the search for their supporters must also be global, and the support of many different groups may be necessary to collect standardised data. This is likely to be very resource intensive, and unlikely to be financially viable without the highest levels of political support.

The collection of data from the illustrative field of biotechnology R&D has revealed difficulty in codifying distinctions between fields of technologies, inability to reflect new and emerging technologies, considerable methodological differences between different institutions and countries in the collection and aggregation of data and the decay of data over time. While there is a demand to have good data on public and private sector investment in new technologies, R&D financing data gathered for policymaking are often problematic, partitial, or contested (House of Lords 2010; Anon 2010; Arundel, van Beuzekom, and Gillespie 2007).

The challenges of collecting 'input' data for innovation (in the form of R&D spending) may be contrasted with the use of proxy, 'output' measures of innovation. Proxies do not directly capture funding but represent important intermediate outputs of innovation processes (Patel 2012). We discuss how centrally gathered and indexed data, based on submissions by organisations for purposes other than being monitored, can be analysed in diverse ways for up to date or historical studies. In addition, these measures have other benefits as well: patent and publication databases are widely available at relatively low cost, and the data on these are carefully indexed over long periods of time (although there are differences over time in organisational activity and coverage that require adjustment). Proxies can also be retrospectively interrogated using different keywords or classifications and without being subject to the sorts of recall bias individuals and organisations suffer from.

While proxy measures have considerable power to document the evolution of technologies, in more detail, with greater scope, less direct reliance on the organisations being observed and at lower cost than similar examinations using R&D data, they also have important drawbacks. The externally validated categories used for proxy measures reflect human judgement and failure to account for this may result in misleading understandings of technologies (Benner and Waldfogel 2008). In addition, particularly for publications the key metrics of citations backward and forward may be subject to considerable technical and game-playing problems (Moed 2002; Aksnes 2003). When applied against our framework we find that both patents and publications present considerably more detail about R&D activities related to emerging technologies than do funding statistics. Our discussion of the genomics sector in Section 6 demonstrates the extent to which funding figures neglect the nuances of innovative activity or even fails to find it entirely.

Additionally, we point to the emergence of tech mining as a tool for analysis of new technologies as an exciting new frontier for the use of information technology, but a prerequisite for advanced analysis is suitable data. We suggest that tech mining techniques may allow researchers to mitigate some of the logistical weaknesses inherent in proxy data by more extensive and sophisticated textual analysis and searching allowing the evolution of ideas and technologies to be studied more easily and at scale.

The intent of this framework and discussion is not intended to suggest that R&D funding statistics have no use or that bibliometric and tech mining techniques are unequivocally superior. Both input and output measures have a role to play in policy discussions and strategic planning, and both may be used together in a complementary fashion in well-resourced dedicated initiatives. However we would suggest that the framework introduced in this paper provides a tool for analysing and understanding the relative strengths and weaknesses of methods of observing emerging technologies. The example of biotechnology R&D discussed here is a particular instance where conflicting categories and difficulties in tracking funding make systematic analysis of inputs to innovation difficult.

An important point on the use of proxies is that the availability of robust datasets relatively free from manipulation by R&D active organisations themselves, varies greatly by technological field and industrial sector. Whereas in the analysis of biotechnology, we benefit from a wide range of proxies as discussed in this paper, analysis of other sectors using advanced techniques may require alternate measures of innovative activity (such as *inter alia*, new products launched or commercial alliances.). The framework we have elucidated here provides to researchers and policymakers a straightforward tool for making judgements about the validity of proxies for active involvement in emerging technological fields. While no single measure is likely to capture the complex picture of activity stemming from investments in emerging technologies, this framework allows the assessment of different measures of such activity, facilitating the relative judgement of suitability and potentially influencing modes of data collection and analysis.

There are a number of areas in which this framework can be applied. This paper as focused on the emergence of biotechnologies, but other technological areas may be addressed. Another potential application would be in the evaluation of measures of innovation in developed and emerging economies. Historically the measurement of capability development in developing countries has been plagued with methodological problems (see Archibuigi and Coco 2004), as some proxy measures of innovation such as patenting and publications are difficult to measure widely outside of the institutions in the developed countries. While we cannot provide a one-size-fits-all solution to this significant problem, our framework provides a meaningful tool for the evaluation of different measures of innovation in these circumstances. The institutional challenges that policymakers face in these circumstances (for instance data sharing) are reflected in our framework, and as such the framework allows comparisons of the effectiveness of a range of measures in capturing innovative activity.

8. Conclusions

Emerging technologies promise much for firms and economies, but the data necessary to track these technologies in order to make well informed policy and strategy are often not available. Where data gathering is undertaken, the purpose determines who collects the data, and who contributes, how much care they take, the form in which data is sought, where it is found, how it is counted or excluded, aggregated or partitioned, unitised, coded, tabulated, stored, analysed and presented. Each step further separates the ultimate reader from the phenomena being observed –

which change as a result of being observed – particularly where funding is involved. Consequently this begs the question of whether it is indeed possible to truly identify a representative picture of investments in a given emergent technology, and even if so, whether it is a good use of researchers' time and effort or the funder's scare resources.

This paper helps to address this question by presenting a framework for examining the strengths and weaknesses of different ways of tracking organisations' activity around emerging technologies. This framework shows that the challenges of systematically identifying all potential funding sources for emerging technologies are fundamentally fraught with difficulties with at least 16 barriers to the collection and analysis of comprehensive funding data. We use illustrative data from biotechnology R&D to argue that proxy measures such as patents or publications can potentially provide more detailed measures of innovation activity in some cases. Finally we suggest that while gathering financial data will remain difficult in most circumstances, tech mining approaches have the potential to mitigate some problems inherent in proxy-based approaches and offer further promise of improvements in the future.

Acknowledgements

This article was supported by a grant from the Economic and Social Research Council (RES-360-25-0076). The article in its current form has been developed from a discussion paper originally commissioned by the Nuffield Council on Bioethics as part of their 2012 report 'Emerging biotechnologies: technology, choice and the public good'. We are grateful to Andy Stirling who first highlighted to us the issue the paper addresses and to the members of the Nuffield's working party and secretariat, our interviewees, Denise Chiavetta and two anonymous reviewers for providing helpful comments that have greatly improved the paper. The views expressed, and any remaining errors, remain our own.

Notes

1. Recent schemes include those in the UK by UKCRC, Researchfish Ltd and Research Councils UK research outcomes tracking system, as well as the NIH's RePORT in the USA.
2. See House of Commons Science and Technology Committee's 2010 Report on 'Setting priorities for publicly funded research Setting priorities for publicly funded research', chap. 3, paras 25–29. Available at http://www.publications.parliament.uk/pa/ld200910/ldselect/ldsctech/104/10406.htm (accessed 15 March 2013).
3. This classical taxonomy of classifications is suggested to be a simplification and cannot be regarded as entirely unproblematic (see Lakoff 1987).
4. http://www.oecd-ilibrary.org/science-and-technology/frascati-manual-2002_9789264199040-en.
5. Levels of private investment are very difficult to access from public reports: In OECD (2009) 24 countries provided data on numbers of biotechnology firms in their national sectors, but only 13 provided data on the industrial applications (e.g. health, agriculture, food and beverages) their firms focused on. This is sufficient to account for the sectoral focus of only 11% of expenditure in the 19 countries providing data on private R&D investments. Consequently it is very difficult to draw meaningful conclusions from these results, given that they are incomplete in terms both of nations providing data and the quality of national provided.
6. See for instance the policy for selection to the ISI journal database at free/essays/journal_selection_process/. http://thomsonreuters.com/products_services/science/free/essays/journal_selection_process/.
7. http://www.nytimes.com/1999/08/29/business/the-race-to-cash-in-on-the-genetic-code.html?pagewanted=all.

Notes on contributors

Michael Hopkins is Senior Lecturer at SPRU: Science and Technology Policy Research, University of Sussex. Michael trained in biology before joining SPRU for his MSc and DPhil. Michael has spent 15 years studying the dynamics of biotech innovation around diagnostics and drugs, especially trends in intellectual property protection and financing. His research, which emphasises the use of mixed methods and constructivist theories to reveal the complexities and uncertainties of biomedical innovation has been supported by a wide range of funders including the ESRC, MRC, ESPRC, NESTA,

Cancer Research UK, the European Commission, and the US NSF. Michael has published over 20 peer-reviewed papers in scientific and management journals, and tweets as @biotechpolicyUK.

Josh Siepel is Lecturer in Management at SPRU, University of Sussex. A biochemist by training, he studied science and technology policy at SPRU as a Marshall Scholar. His research interests include the role of institutions in the support of firm growth, and on the uses and implication of quantification for the shaping of policy, particularly in the area of small firm finance. His research includes work for the UK Department of Business Innovation and Skills, NESTA, the European Commission, ESRC and British Venture Capital Association.

References

Abraham, J. 2010. Pharmaceuticalization of society in context. *Sociology* 44: 603–22.

Aksnes, D. 2003. A macro study of self-citation. *Scientometrics* 56, no. 2: 235–46.

Alcacer, J., and M. Gittelman. 2006. Patent citations as a measure of knowledge flows: The influence of examiner citations. *Review of Economics and Statistics* 88: 774–79.

Alcacer, J., M. Gittelman, and B. Sampat. 2009. Applicant and examiner citations in US patents: An overview and analysis. *Research Policy* 38: 415–27.

Anon. 2010. Editorial. Wrong Numbers? *Nature Biotechnology* 28: 761.

Archibuigi, D. 1992. Patenting as an indicator of technological innovation: A review. *Science and Public Policy* 19: 357–68.

Archibuigi, D., and A. Coco. 2004. A new indicator of technological capabilities for developed and developing countries (ArCo). *World Development* 32: 629–54.

Archibuigi, D., and M. Pianta. 1996. Measuring technological change through patents and innovation surveys. *Technovation* 16: 451–68.

Aristotle. 1995. Categories. In *The complete works of Aristotle*, ed. Jonathan Barnes, trans. J.L. Ackrill, 3–24. Princeton, NJ: Princeton University Press.

Arundel, A.V., B. van Beuzekom, and I. Gillespie. 2007. Defining biotechnology – carefully. *Trends in Biotechnology* 25: 331–32.

Augsdorfer, P. 1996. *Forbidden fruit: An analysis of bootlegging, uncertainty and learning in corporate R&D*. Aldershot: Avebury.

Benner, M., and J. Waldfogel. 2008. Close to you? Bias and precision in patent-based measures of technological proximity. *Research Policy* 37: 1556–67.

Bijker, W.E. 1995. *Of bicylces, bakelites and bulbs*. Cambridge, MA: MIT Press.

BIS. 2010. Life sciences in the UK – economic analysis and evidence for life sciences 2010: Delivering the blueprint. BIS economics paper No. 2, Department of Business, Innovation and Skills and Department of Health.

Bornmann, L. 2011. Mimicry in science? *Scientometrics* 86: 173–77.

Bowker, G., and S. Star. 2000. *Sorting things out: Classification and its consequences*. Cambridge, MA: MIT Press.

Boyack, K., R. Klavans, and K. Borner. 2005. Mapping the backbone of science. *Scientometrics* 64: 351–74.

Calvert, J. 2006 .What's special about basic research? *Science Technology and Human Values* 31: 199–220.

Cozzens, S., S. Gatchair, J. Kang, K.-S. Kim, H.J. Lee, G. Ordóñez, and A. Porter. 2010. Emerging technologies: Quantitative identification and measurement. *Technology Analysis & Strategic Management* 22, no. 3: 361–76.

Dosi, G. 1982. Technological paradigms and technological trajectories: A suggested interpretation of the determinants and directions of technical change. *Research Policy* 11: 147–62.

Enzing, C., A. van de Gisessen, S. van der Molen, G. Manicad, T. Reiss, R. Linder, D. Lacasa, et al. 2007. *BioPolis: Inventory and analysis of national public policies that stimulate biotechnology research, its exploitation and commercialisation by industry in Europe in the period 2002–2005*. Brussels: European Commission. http://ec.europa.eu/research/biosociety/pdf/biopolis-finalreport_en.pdf.

Freeman, C., and F. Louca. 2001. *As time goes by*. Oxford: OUP.

Gassmann, O., G. Reepmeyer, and M. von Zedtwitz. 2004. *Leading pharmaceutical innovation: Trends and drivers for growth in the pharmaceutical industry*. London: Springer.

Gittelman, M., and B. Kogut. 2003. Does good science lead to valuable knowledge? Biotechnology firms and the evolutionary logic of citation patterns. *Management Science* 49: 366–82.

Glanzel, W., and H. Moed. 2002. Journal impact measures in bibliometric research. *Scientometrics* 53: 171–93.

Haggerty, K. 2001. Negotiated measures: The institutional micropolitics of official criminal justice statistics. *Studies in History and Philosophy of Science Part A* 32: 705–22.

Hall, B., A. Jaffe, and M. Trajtenberg. 2001. The NBER patent citation data file: Lessons, insights and methodological tools. NBER working paper 8498.

Headrick, D. 2000. *When information came of age*. Oxford: Oxford Univerity Press.

Hess, A., and F. Rothaermel. (2011). When are assets complementary? Star scientists, strategic alliances, and innovation in the pharmaceutical industry. *Strategic Management Journal* 32: 895–909.

Hicks, D. 1995. Published papers, tacit competencies and corporate management of the public/private nature of knowledge. *Industrial and Corporate Change* 4: 401–424.

Hicks, D. 2011. Structural change and industrial classification. *Structural Change and Economic Dynamics* 22: 93–105.

Hojat, M., J. Gonnella, and A. Caelleigh. 2003. Impartial judgment by the 'gatekeepers' of science: Fallibility and accountability in the peer review projecess. *Advances in Health Sciences Education* 8: 75–96.

Hopkins, M., P. Martin, P. Nightingale, A. Kraft, and S. Mahdi. 2007. The Myth of the Biotech Revolution: An assessment of technological, clinical and organizational change. *Research Policy* 36, no. 4: 566–89.

Hopkins, M., S. Madhi, P. Patel, and S. Thomas. 2007. DNA patenting: End of an era? *Nature Biotechnology* 25: 185–87.

Hopkins, M.M., S. Mahdi, S. Thomas, and P. Patel. 2006. Global trends in the granting, filing and exploitation of DNA patents. A report for the European Commission. Brighton: SPRU. http://www.sussex.ac.uk/spru/documents/patgen_finalreport.pdf (accessed October 2008).

House of Lords. 2010. *Science and Technology Committee Examination of Winesses Questions* 436–39. http://www.publications.parliament.uk/pa/ld200910/ldselect/ldsctech/104/10011208.htm

Huggett, B., J. Hodgson, and R. Lähteenmäki. 2011. Public biotech 2010 – the numbers. *Nature Biotechnology* 29: 585–91.

Jacobsson, S., and A. Johnson. 2000. The diffusion of renewable energy technology: An analytical framework and key issues for research. *Energy Policy* 28: 625–40.

Jaffe, A.B., and M. Trajtenberg. 2002. *Patents, citations, and innovations: A window on the knowledge economy*. Cambridge, MA: MIT Press.

Jaffe, A., M. Trajtenberg, and R. Henderson. 1993. Geographic evidence of knowledge spillovers as evidenced by patent citations. *Quarterly Journal of Economics* 108, no. 3: 577–98.

Lakoff, G. 1987. *Women, fire and dangerous things*. Chicago, IL: Chicago University Press.

Larsen, P., and M. von Ins. 2010. The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. *Scientometrics* 84: 575–603.

Leydesdorff, L. 1987. Various methods for the mapping of science. *Scientometrics* 11, nos. 5–6: 295–324.

Leydesdorff, L., D. Rotolo, and I. Rafols. 2012. Bibliometric perspectives on medical innovation using the medical subject headings of PubMed. *Journal of the American Society for Information Science and Technology* 63, no. 11: 2239–53.

Martin, B. (1996). The use of multiple indicators in the assessment of basic research. *Scientometrics* 36, no. 3: 343–62.

Martin, P., M.M. Hopkins, P. Nightingale, and A. Kraft. (2009). On a critical path: Genomics, the crisis of pharmaceutical productivity and the search for sustainability. In *Hand book of genetics and society*, ed. P. Atkinson, P. Glasner, and M. Lock, 145–62. Abingdon, UK: Routledge.

Matthews, K., and M. Rowland. 2011. *Stem cells and biomedical research in Texas*. Houston, TX: Science & Technology Policy, James A. Baker III Institute for Public Policy, Rice University.

Meyer, M. 2000. What is special about patent citations? The differences between scientific and patent citation. *Scientometrics* 49, no. 1: 93–123.

Moed, H. 2002. The impact factors debate: The ISI's uses and limits. *Nature* 415: 731–32.

Morgan Jones, M., and J. Grant. 2011. *Complex trauma research in the UK*. Cambridge, UK: RAND Europe.

Morillo, F., M. Bordons, and I. Gomez. 2003. Interdisciplinarity in science: A tentative typology of disciplines and research areas. *Journal of the American Society for Information Science and Technology* 54: 1237–49.

Murray, F. 2002. Innovation as co-evolution of scientific and technical networks: Exploring tissue engineering. *Research Policy* 31, nos. 8–9: 1389–407.

NESTA. 2006. *The innovation gap.* London: NESTA.

Nicholaisen, J. 2008. Citation analysis. *Annual Review of Information Science and Technology* 41: 409–41.

Nuffield Council on Bioethics (2012). *Emerging biotechnologies: Full report*. London: Nuffield Council on Bioethics.

OECD. 2005. *A framework for biotechnology statistics*. Paris: OECD.

OECD. 2009. *Biotechnology statistics 2009*. Paris: OECD.

Patel, P. 2012. *Indicators of innovation performance*. From Knowledge Management to Strategic Competence, Series on Technology Management 19. London: Imperial College Press, 137–95.

Pavitt, K. 1985. Patent statistics as indicators. *Scientometrics* 7, no. 1–2, 77–99.

Porter, T. 1992. Quantification and the accounting ideal in science. *Social Studies of Science* 22: 633–51.

Porter, T. 1996. *Trust in numbers: The pursuit of objectivity in science and public life*. Princeton, NJ: Princeton University Press.

Power, M. 2004. Counting, control and calculation: Reflections on measuring and management. *Human Relations* 57: 765–83.

Power, M. 2005. The invention of operational risk. *Review of International Political Economy* 12: 577–99.

Pudovkin, A., and E. Garfield. 2002. Algorithmic procedure for finding semantically related journals. *Journal of the American Society for Informaiton Science and Technology* 53: 1113–19.

Rafols, I., M. Hopkins, J. Hoekman, J. Siepel, A. O'Hare, A. Perianes-Rodriguez, and P. Nightingale 2012. Big pharma, little science? A bibliometric perspective on big pharma's R&D decline. *Technological Forecasting and Social Change*. http://dx.doi.org/10.1016/j.techfore.2012.06.007

Rosenberg, N. 1982. *Inside the black box: Technology and economics*. Cambridge: Cambridge University Press.

Rothman, H., and A. Kraft 2006. Downstream and into deep biology: Evolving business models in 'top tier' genomics companies. *Journal of Commercial Biotechnology* 12, no. 2: 86–98.

Searle, J. 1995. *The construction of social reality*. London: Allen Lane.

Sismondo, S. 2007. Ghost management: How much of the medical literature is shaped behind the scenes by the pharmaceutical industry? *PLoS Medicine* 4, e286.

Sismondo, S. 2009. Ghosts in the machine. *Social Studies of Science* 39: 171–98.

Star, S., and J. Griesemer. 1989. Institutional ecology, 'translations' and boundary objects: Amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907–93. *Social Studies of Science* 19: 387–420.

Smith, A., J. Voβ, and J. Grin. 2010. Innovation studies and sustainability transitions: The allure of the multi-level perspective, and its challenges. *Research Policy* 39: 435–48.

Strumsky, D., J. Lobo, and S. vander Leeuw. 2012. Using patent technology codes to study technological change. *Economics of Innovation and New Technology* 21, no. 3: 267–86.

Tan, D., and P. Roberts. 2010. Categorical coherence, classification volatility and examiner-added citations. *Research Policy* 39: 89–102.

Thomson Reuters 2012. The Thomson Reuters Journal Selection Process. http://thomsonreuters.com/products_services/science/free/essays/journal_selection_process/ (Revised May 2012).

van Leeuwen, T., H. Moed, R. Tijssen, M. Visser, and A. Van Raan. 2001. Language biases in the coverage of the science citation index and its consequences for international comparisons of national research performance. *Scientometrics* 51, no. 1: 335–46.

Verbong, G., and F. Geels. 2007. The ongoing energy transition: Lessons from a socio-technical, multi-level analysis of the Dutch electricity system (1960–2004). *Energy Policy* 35: 1025–37.

Veugelers, R. 2005. An economists' view on bibliometrically measuring scientific research. *Measurement* 3, no. 1: 33–37.

Williams, S.J., P. Martin, and J. Gabe. 2011. The pharmaceuticalisation of society? A framework for analysis. *Sociology of Health and Illness* 33: 710–25.

Yang, Y., L. Akers, T. Klose, and C.B. Yang. 2008. Text mining and visualization tools–impressions of emerging capabilities. *World Patent Information* 30: 280–93.